A GENERAL MODEL OF REGRESSION USING ITERATIVE SERIES

NILOTPAL KANTI SINHA

Abstract. We present a new and general method of weighted least square univariate regression where the dependent variable is expanded as a series of suitably chosen functions of the independent variables. Each term of the series is obtained by an iterative process which reduces the sum of the square of the residuals. Thus by evaluating the regression series to a sufficiently large number of terms we can, in principle, reduce the sum of the square of residuals and improve the accuracy of the fit.

1. Introduction

In the traditional models of regression, the relationship between the predicted variable y and the predictor variable x is expressed as

$$y_i = f(\beta, x_i) + \epsilon_i \tag{1.1}$$

where the function f is not completely known but is known up to s set of parameters $\beta = (\beta_1, \beta_2, \beta_3, \ldots)$ and ϵ_i is the error term of observation i. The primary problem in the development of a statistical theory and application of statistical methodology is the selection of a suitable model which is a formalization of the relationship between variables in the form of mathematical equations.

The existing methods of regression have the following limitations. First, a given dataset fits into a model in only one way and therefore the as soon as a model is chosen, SS gets automatically fixed. Since the parameters are optimized to obtain the curve of best fit, there is no scope for the user to reduce SS further unless a different model is chosen. Second, no single function f fits the model 1.1 for all datasets sufficiently accurately. A function which is suitable for a particular dataset may be unsuitable for another dataset; e.g. a linear regression is suitable for data that is roughly linear but for highly non-linear data, using linear regression could lead to inaccurate analysis.

Consider the analogy of functions $f(x)$ which satisfy the conditions of Taylor's theorem and can be expanded as a general Taylor series in terms of x. In the Taylor expansion of any $f(x)$ the concept of power series expansion is common to all functions f; only the coefficients vary across the functions. Therefore in terms of models, we can say that Taylor expansions are a family of models that will fit all functions which satisfy the conditions of Taylor's theorem.

The functions on which Taylor's theorem can be applied are continuous but the datasets on which regression is performed are discrete. This brings us to the question whether we can formulate a discrete analogy of Taylor's expansion. In other words, does there exist a general regression model that will give a sufficiently good fit for all datasets? In particular, can we have a method of regression with all the following features embedded in the same model?

Key words. Regression, iterations, sinusoidal series.
2000 *Mathematics Subject Classification.* 93E24, 62J05.

- The model should fit all types of data such as linear, polynomial, data with no visible trend.
- The model should fit seasonal data and capture the periodic patterns as in a time series.
- The user should be able to choose the values of such as Chi-square or R^2.
- The model should not suffer from the problem of over fitting.

In this paper, we answer the question raised above in affirmative and develop the theoretical concepts for a general family of models that will describe all datasets. In particular We consider a univariate response y that we shall relate to a (possibly multivariate) predictor variable x. We shall first develop the concept for the case of a univariate and then extend the theory to the case of multivariate predictor variables.

2. Iterative approach to regression

Let $f_0(\beta, x_i)$ be any model that approximates a given set of n data points with variable $(x_i, y_i), (i = 1, 2, \ldots, n)$. The curve $f_0(\beta, x_i)$ is not necessarily the curve of best fit. Let w_i be the weights assigned to the corresponding sum of the squares of the residuals. We assume that $w_i > 0$. The weighted sum of the square of the residuals is

$$\sum_{i=1}^{n} w_i \{y_i - f_0(\beta, x_i)\}^2.$$

Without loss of generality, we assume that not all x_i are zeroes. Let $f_1(\beta, x_i)$ be a function such that

$$\sum_{i=1}^{n} w_i \{y_i - f_0(\beta, x_i) - t f_1(\beta, x_i)\}^2 < \sum_{i=1}^{n} w_i \{y_i - f_0(\beta, x_i)\}^2 \tag{2.1}$$

so that

$$y = f_0(\beta, x) + t f_1(\beta, x) \tag{2.2}$$

is a model with a lesser sum of the square of the residuals than the model $y_i = f_0(\beta, x_i)$. Our objective is to find the optimal value of t which will minimize the L.H.S. in 2.1.

2.1. Reducing the sum of the square of the residuals using the point of minima of a quadratic equation.

Simplifying 2.1, we obtain the quadratic equation in t

$$E = t^2 \sum_{i=1}^{n} w_i \{f_1(\beta, x_i)\}^2 - 2t \sum_{i=1}^{n} w_i \{y_i - f_0(\beta, x_i)\} f_1(\beta, x_i) < 0.$$

$$\frac{dE}{dt} = 2t \sum_{i=1}^{n} w_i \{f_1(\beta, x_i)\}^2 - 2 \sum_{i=1}^{n} w_i \{y_i - f_0(\beta, x_i)\} f_1(\beta, x_i) = 0$$

or

$$\alpha_k = \frac{\sum_{i=1}^{n} w_i \{y_i - f_0(\beta, x_i)\} f_1(\beta, x_i)}{\sum_{i=1}^{n} w_i \{f_1(\beta, x_i)\}^2}. \tag{2.3}$$

Also

$$\frac{d^2E}{dt^2} = 2 \sum_{i=1}^{n} w_i \{f_1(\beta, x_i)\}^2 > 0.$$

Since the second derivative is positive, E has a minima at the value of t given by 2.3. Hence this is the optimal value of t at which the L.H.S. of 2.1 will be minimum. This gives us a method to reduce the sum of the square of the residuals.

We can repeat the above process of reducing the sum of the square of the residuals by replacing $f_0(\beta, x)$ with $f_0(\beta, x) + \alpha_1 \beta, x)$, where α_1 is the optimal value of t obtained in the first iteration. Hence by successive iteration we obtain a series of the form

$$y_i = f_0(\beta, x_i) + \alpha_1 f_1(\beta, x_i) + \alpha_2 f_2(\beta, x_i) + \dots$$

After each iteration we calculate SS. We stop the iterations after SS has shrunk below a maximum acceptable value. Theoretically, if each x_i is unique then we can have $SS \to 0$ by iterating a sufficiently large number of times. However this can lead to over fitting which occurs when a statistical model is complex and has too many parameters relative to the number of observations. An over fitted model is trained to describe random error instead of the underlying relationship.

Notice that for a given value x our regression model will give only one value of y and this is true for all other models of the form 1.1 where x and f one-one-relationship. In real life data, it is possible that there are repetitions of the independent variable x which give two or more distinct values of the dependent variable y. In such a scenario the limiting value of the sum of the square of the residuals will not approach zero. For example if (x, a) and (x, b) are two distinct values of y for the same value of x in a survey, and every other value of x is different in the collected sample then applying our method of regression, the limiting value of the sum of the square of the residuals will be $(a - b)^2/2$.

In such a scenario, in order to build a regression fit where $SS \to 0$, we can consider only one of the sample point from the repeated observations of independent variable x or consider a new point whose dependent variable is the mean value of the dependent values of all the repeated independent variables x. This is consistent with the assumption of regression that the independent variables are uncorrelated.

2.2. **Choice of regression functions.** We shall call f_0 as the initial approximation or base model and $f_1, f_2, \dots$ as the regression functions. It is desirable to choose suitable $f_0, f_1, f_2, \dots$ and the parameters $\beta = (\beta_1, \beta_2, \beta_3, \dots)$ so as to accelerate the rate of convergence of SS. One of the advantage of our method is that the choice of the initial approximation and the regression functions is with the user and therefore these function can be chosen based on the dataset under study. If the dataset shows a trend, say linear or polynomial or any other form that can be determined with a preliminary regression, then we can use that fit. If however the dataset is completely erratic and shows no particular trend or if the trend cannot be determined, we can take $f_0(\beta, x)$ to be a constant. A good starting value of this constant is the mean of the dependent variables. If we want a model that is independent of the constant term, we can take $f_0(\beta, x)$.

3. Motivation: Fourier series analogy for discrete points

In our investigation on regression functions f, we found that sinusoidal functions of the form $\sin(\beta g(x))$ to be suitable. Here $g(x)$ is an arbitrary function used to control the sensitivity of dependent variables to small variations in the independent variables. In this section, we lay down the steps for sinusoidal regression method. The motivations for studying the sinusoidal functions is as follows:

- **Its resemblance to Fourier series**. A Fourier series is an expansion of a periodic function $f(x)$ in terms of an infinite sum of sines and cosines. A sinusoidal regression will be a Fourier series analogy for discrete points.

4. The method of iterative regression

The following are the steps for iterative regression using sinusoidal series.

1. Choose an initial approximation $f_1(x)$. Let

$$\alpha_1 = \frac{\sum_{i=1}^{n} w_i\{y_i - f_1(x_i)\}\sin(\beta_i g(x_i)}{\sum_{i=1}^{n} w_i \sin^2(\beta_1 g(x_i))}.$$

2. The new estimate of regression fit is $y = f_1(x) + \alpha_1 \sin(\beta_i g(x_i))$. It is desirable to find the optimal β_1 which minimizes SS.

3. Take $y = f_1(x) + \alpha_1 \sin(\beta_i g(x_i))$ as the new initial approximation and repeat step 2. Continue this iterative process until the convergence criteria imposed on SS is satisfied (say after m iterations). The required regression curve is

$$y = f_1(x) + \sum_{j=1}^{m} \alpha_j \sin(\beta_j x). \tag{4.1}$$

Theorem 4.1. *Every set of finite set of discrete points (x_i, y_i) where x_i is unique can be expanded as sinusoidal series.*

Proof. The proof follows from the fact that since each x_i is unique, the limiting value of SS will be zero. □

5. An application: Periodicity in solar eclipses across centuries

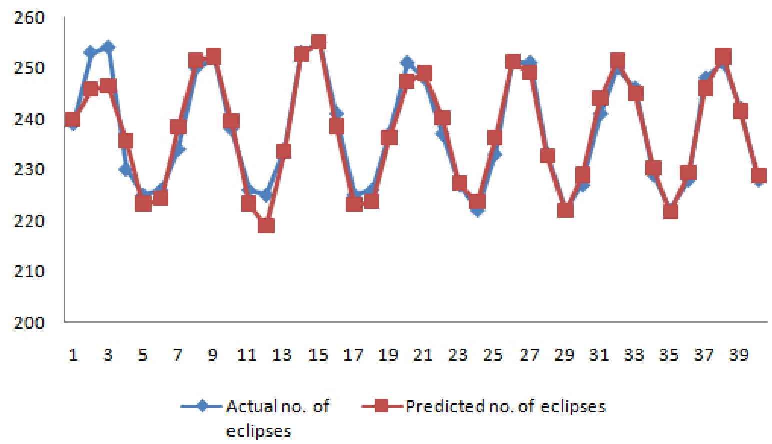
For all practical purposes, the Sun, the Earth and the Moon can be considered to be a stable system with deterministic positions. Therefore the number of solar eclipses in given time interval should depend only on the length of the interval. We shall apply sinusoidal regression to the number of solar eclipses in a time interval and unearth a near periodic pattern in the occurrence of solar eclipses. Since solar eclipses are rare, we need large time intervals that contains sufficient numbers of solar eclipses to enable us to perform statistical analysis. NASA has published the data for the total number of solar eclipses in a century from 19^{th} century BC to 30^{th} AD (See [6]). Let $E(n)$ denote the total number of solar eclipses in the n^{th}. Since n is unique, we can obtain a sinusoidal regression of the form.

$$E(n) = E_0 + \sum_{i=1}^{\infty} \alpha_i \sin(\beta_i n) \tag{5.1}$$

where E_0 is a suitably chosen constant. The total number of solar eclipses in a century varied between 222 and 256 therefore we expect the total number of solar eclipses in a century to be close to this range. Hence the initial approximation should be a function that is unbounded at $\pm\infty$. The simplest function satisfying this condition is the constant function. This justifies the choice of E_0 as a constant. The actual value of E_0 not important as it acts as a scaling factor and the rest of the parameters would adjust accordingly to a chosen value of E_0. Using the data from the 19^{th} century BC to 20^{th} AD we obtain the sinusoidal curve (with parameters rounded off to two decimal places)

$$E(n-20) = 237.23 + 11.02\sin(n) - 8.33\sin(1.14n) + 4.58\sin(0.88n)$$

$$-2.20\sin(1.31n) - 1.81\sin(1.61n) + 1.53\sin(1.07n). \tag{5.2}$$



The $(n-20)$ in the LHS is again a scaling adjustment since k^{th} BC was taken as $-k$ while computing values of the parameters. Irrespective of the choice of the scaling factors and shift of origin, we can always fit a general model of the form 5.2 to the number of eclipses in a century.

The above graph shows the plot of the actual number of eclipses and the number given by the sinusoidal model. In only six iterations we have reduced SS from 4840.44 to 301.62. This reduction in SS gives a very good fit as shown in the above graph which plots the actual number of eclipses and the number of eclipses given by the model.

Since the parameters in 5.2 have been rounded off to two decimal places, 5.2 has an approximate period of 200π. However the total number of eclipses in a century is a natural number; therefore considering only the integer part of $E(n-20)$, we observe that it has a quasiperiod of about 2π centuries which in this case corresponds to a time interval of about six centuries. Based on this empirical evidence we formulate the following hypothesis (which most probably is already known to the astronomers).

Hypothesis: *The number of solar eclipses in a century roughly repeats every sixth century.*

How consistent is this hypothesis with actual data? NASAs solar eclipse data tells us that the number of solar eclipses in the centuries 18^{th} BC, 12^{th} BC, ... , 12^{th} AD, 18^{th} AD are 253, 250, 253, 251, 251, 250 and 251 respectively. Similarly in the 15^{th} BC, 9^{th} BC, ... , 15^{th} BC and 21^{th} BC centuries, the total number of solar eclipses are 225, 226, 225, 227, 222, 222, 224. We have just discovered a beautiful law of nature. Surely astronomers would have already found this using the equations of gravity; nonetheless we have discovered this on our own using the power of iterative regression. This would not have been possible had we used traditional regression.

6. Conclusions and scope of future works

A model is typically trained by maximizing its performance on some set of training data. However the efficacy of a model is determined not by its performance on the training data but by its ability to perform well on unseen data. An over fitted model will typically fail drastically on unseen data and the value of SS will shrink relative to the original training data. Therefore

for practical applications of our method of regression, a balance between minimizing SS and the number of iterations and, the choice of regression functions is necessary.

No mathematical model based on past data can accurately predict the future. However better mathematical models help in reducing risk and this is where the flexibility of in choosing the initial approximation and regression functions in our method of regression can have an edge over the traditional methods. For example, we can find mathematical functions that roughly describes observed phenomenon in scientific or business application and then use these functions in our iterative process to obtain regression fit with improved accuracy. Developing such application based functions will be of immense value in forecasting and prediction. Our new method of regression opens up vast scope for future research, some of which have been listed below.

Finally we would like to develop iterative regression for multivariate relationship. The author is already working on this and this will be the topic of future paper.

7. Avoiding Overfitting

Overfitting is introduced whenever we over optimise a performance measurement criterion based on a finite sample of data, resulting in a model which is excessively complex, such as having too many parameters relative to the number of observations and as a results, the model ends up describing random error or noise present in the data instead of the underlying relationship that produced the data. Such models have poor predictive performance, as it can exaggerate minor fluctuations in the data.

In iterative regression, we run the risk of continuing the iterations beyond an optimal number of times. To avoid this problem, the standard approaches to prevent overfitting can be employed, especially the method of early stopping. In this method, the training set is further split into a training set and a validation set. After each iteration through the new training set, the model is evaluated on the validation set. When the performance with the validation test stops improving, the iteration is stopped.

The method of iterative regression will eventually start describing the randon errors or noise if the iteration is not early stopped. Hence it is essential to ensure that the training data set have been observed under identical conditions and all outliers have been removed.

8. Acknowledgement

The author is grateful to Dr. V Srinivasan, Adams Distinguished Professor of Management, Emeritus, Stanford Business School, USA; Dr. Purba Rao, Asian Institute of Management, Phillipines; and Dr. Dr. S Bharadwaj, Great Lakes Institute of Management for their valuable guidance and suggestions.

References

[1] A.M. Legendre. Nouvelles mthodes pour la dtermination des orbites des comtes (1805). Sur la Mthode des moindres quarrs appears as an appendix.
[2] C.F. Gauss, Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum. (1809)
[3] C.F. Gauss, Theoria combinationis observationum erroribus minimis obnoxiae. (1821/1823)
[4] Christian Ritz and Jens Carl, Streibig, Nonlinear regression with R, Springer Publications, pp - 55.
[5] Helmut Spaeth, Mathematical Algorithms for Linear Regression, Academic Press, ISBN 0-12-656460-4, page 304 (1991).
[6] http://eclipse.gsfc.nasa.gov/SEcat5/SEcatalog.html
[7] D. Kundu, Asymptotic properties of the least squares estimators of sinusoidal signals, Statistics, 30, 221-238 (1997)

[8] D. Kundu, Estimating parameters of sinusoidal frequency; some recent developments", National Academy of Sciences Letters, vol. 25, 53-73 (2002).

[9] D. Brillinger, Fitting cosines: Some procedures and some physical examples", Applied Probability, Stochastic Process and Sampling Theory, Ed. I.B. MacNeill and G. J. Umphrey, 75-100, Dordrecht: Reide (1987).

[10] C. T. Kelley, Iterative Methods for Optimization, SIAM Frontiers in Applied Mathematics, no 18, ISBN: 0-89871-433-8 (1999).

[11] Å. Björck, Numerical Methods for Least Squares Problems. SIAM. ISBN 978-0-898713-60-2 (1996).

Fractal Analytics, Mumbai, Maharastra, 400063, India
E-mail address: nilotpals@fractalanalytics.com, nilotpalsinha@gmail.com